Fast and Robust Moving Objects Detection based on Non-parametric Background Modeling

Jianping Han
Institute of Graphics and Image, Hangzhou Dianzi University, Hangzhou, China
Email: hanjp@hdu.edu.cn

Zhigeng Pan, Mingmin Zhang
State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China
Email: {zgpan, zmm }@cad.zju.edu.cn

***Abstract*—Fast and reliable detection of moving objects is one of the important requirements for many computer vision and video analysis applications. Mean shift based non-parametric background modeling supports more sensitive and robust detection in dynamic outdoor scenes. However it is prohibitive to real-time applications such as video surveillance. This paper aims to deal with the limitation of high computational complexity. Firstly, coarse to fine methods are proposed to avoid raster scanning entire image. Foreground pixels are detected in coarse level to roughly locate the foreground objects in the image, and then fine detection is performed on the corresponding blocks gradually. Secondly， fast mean shift approach is presented according to temporal dependencies. Mean shift iterations are performed starting from incoming data and the modes obtained last time. The experimental results show that the proposed algorithm is effective and efficient in dynamic environment. The proposed algorithm has been applied to move objects detection in our real-time marine video surveillance system.**

***Index Terms*—background subtraction, nonparametric, mean shift**

I. INTRODUCTION

Detection and segmentation of moving objects in video sequence is a basic task in many computer vision and video analysis applications, such as video surveillance, indexing for multimedia, and perceptual human-computer interface. A mostly used approach is background subtraction, where each video frame is compared with the background model extracted from previous frames, pixels that deviate significantly from the background are considered to be moving objects. The critical step of this kind of methods is how to build and maintain the background model. Despite of its importance, the task is still far from being completely solved since the backgrounds are usually dynamic in nature such as swaying leaves, ripple water, etc. A robust background modeling technique should also handle situations in which new objects are introduced to or old ones are removed from the background. Further more, the background modeling algorithm should be executed in real-time.

Corresponding Author: Zhigeng Pan(zgpan@cad.zju.edu.cn)

A. Previous work

Many methods are presented to build and maintain the background model for background subtraction. One very popular technique that models each pixel with a Gaussian distribution [4] does not work well in the case of dynamic natural scenes. In ref. [5], each pixel is modeled as a mixture of fixed number of Gaussians (MoG). In MoG, an online K-means approximation is used instead of the exact Expectation Maximization algorithm. Methods employing MoG have been widely incorporated into algorithms that utilize color and gradient information [13], Bayesian framework [6], and region based information [7]. Yuting Chen [15] proposes efficient hierarchical method, in which rough foreground objects are detected in coarse block level in advance. Then based on the result of coarse block level, fine pixel-level method is performed to further extract the detailed shapes of foreground objects.

Limitation of MoG is that it is necessary to specify the number of Gaussians, and generally speaking, it is difficult to add or remove components in a principled way. Therefore, this technique may not be flexible enough to background with fast variations. Nonparametric method is initially proposed by Elgammal [1] for more general background modeling. The probability density function for pixel intensity is estimated directly from the data without any assumptions about the underlying distributions. Sheikh and Shah [18] unify the temporal and spatial consistencies into the nonparametric model. Similar models include [2, 14]. As an elegant way to locate the modes of the intensity value distribution, mean shift has been applied to nonparametric background modeling. In [17], mean shift procedure is used for locating the most reliable background mode. In [9], Wang adapts the kernel used in mean shift to be anisotropic and achieves better results in segmenting video. Due to its iterative nature, the computational cost of mean shift procedure is generally high and prohibitive in real-time applications such as

doi:10.4304/jsw.4.10.1084-1090

video surveillance. Many efforts have been devoted to reduce the computational cost. Kai Zhang et al [11] adopt the idea of nearest neighbor consistency in mean shift, and develop a fast mean shift algorithm that significantly reduces the complexity of feature space analysis. M.Piccurdi [12] introduced the concept of local basins of attraction and histogram-based mean shift computation. However, having to perform on each channel seperately, the mean shift computation in [12] does not work well for mode detection in multiple channel images. Ke et al proposed [16] temporal mean shift algorithm via using the cluster locations found in the previous frame to initialize the mode search for the next frame, and achieved a speedup of nearly 100%. In [3], mean shift mode detection from samples is used at initialization time, and it will be replaced by the real-time model update which is provided by simple heuristics coping with mode adaptation, creation, and merging.

B. Overview of our method

In this paper, we present mean shift based non-parametric background modeling for sensitive and robust detection of moving objects in dynamic scenes. To deal with the limitation of high computational complexity of mean shift procedure, some computational optimizations are proposed. Firstly, coarse to fine method is used to avoid the raster scanning of entire image. Foreground pixels are detected in coarse level to roughly locate the foreground objects in the image, and then fine detection is performed on the corresponding blocks gradually. Secondly, to reduce the computational cost of background identification of a pixel, mean shift iterations are performed, starting from incoming data and the mode found last time instead of the whole data set. Our method is utilized to match the capabilities of nonparametric background models in order to model dynamic backgrounds effectively but with greatly reduced computational complexity.

The rest of this paper is organized as follow. In the next section, we describe how mean shift can be used for background modes detection. In Section 3, coarse to fine background subtraction method is presented. In Section 4, fast mean shift procedure is proposed for background modes detection on sampled pixels. Experimental results are presented in Section 5 and Section 6 concludes this paper.

II. MEAN SHIFT FOR BACKGROUND MODES DETECTION

The background modeling and subtraction technique generally assumes that the background is visible more frequently than any foregrounds, so we can relate the estimate of the background to the modes (local maxima) of the underlying distribution. The intensity values of a static pixel can be modeled by a single mode distribution, while dynamic pixels are showed as multi-mode. The mean shift technique is an iterative gradient ascent method with excellent convergence properties which allows it to detect the main modes of a distribution.

Let $\{x_i\}_{i=1,2,\ldots,N}$ be a sample of intensity values for a pixel along the time axis and N is the sample size. Given this sample, we can obtain an estimate probability density function of pixel intensity at any intensity value using kernel density estimator as follow.

$$\hat{f}(x) = \frac{1}{N(2\pi)^{3/2}} \sum_{i=1}^{N} |H|^{-1/2} \exp(-\frac{1}{2} D^2(x, x_i; H)) \quad (1)$$

Where H is bandwidth matrix, Gaussian function is used as the kernel. $D^2(x, x_i; H) = (x - x_i)^T H^{-1} (x - x_i)$ is Mahalanobis distance between x and x_i. Taking the gradient of Equation (1), we can obtain mean shift vector as Equation (2).

$$m(x) = \frac{\sum_{i=1}^{N} x_i exp(-\frac{1}{2} D^2(x, x_i; H))}{\sum_{i=1}^{N} exp(-\frac{1}{2} D^2(x, x_i; H))} - x \quad (2)$$

$$x_j^{(k+1)} = m(x_j^{(k)}) + x_j^{(k)} \quad (3)$$

For each x_j, we can start the iteration as (3) with $x_j^{(0)} = x_j$, and let $(\mu_i, w_i)_{i=1,\ldots,q}$ be the weighted modes, where μ_i is the intensity value. The number of points for each mode is denoted by l_i, then the weight of each mode w_i *can* be defined as

$$w_i = l_i / \sum_{i=1}^{q} l_i \quad (4)$$

Weighted modes are ordered by the value of w_i. Then the first B Gaussians are chosen as the background models, where B is defined as Equation (5). where T is a measure of the minimum portion of the data for which should be accounted by the background.

$$B = \min_b (\sum_{k=1}^{b} \omega_k > T) \quad (5)$$

Mean shift based non-parametric background modeling supports more sensitive and robust detection in dynamic scenes[12]. However, standard implementation of the mean shift is not possible due to the excessive computational consumption. To deal with the limitation, in follow subsection we will propose a procedure which is able to limit the computational consumption drastically.

III. COARSE TO FINE BACKGROUND SUBTRACTION

Rather than isolated pixels, moving objects are generally small regions in a video frame. Therefore there is no need to raster scan entire image to test all pixels against the corresponding background model which aims to detect foreground. Instead, we use Coarse to fine method to sample small portion of the pixels, and increase the efficiency of the mean shift based background subtraction algorithm. Via using a regularly decomposed region quadtree[10], we detect foreground pixels in coarse level to locate the foreground objects in a frame roughly, and perform fine detection on the corresponding blocks gradually.

A pixel-based background subtraction can be generally characterized as a quadruple $\{F,M(t),\Phi, \Gamma\}$, where F depicts the feature extracted for a pixel, $M(t)$ consists of the background models recorded at time t for the pixel, Φ is a function determining whether a given pixel q at time t is background based on the pixel feature, and Γ is another function that updates the model and generates a new model at time $t+1$ based on the pixel feature and the current model $M(t)$.

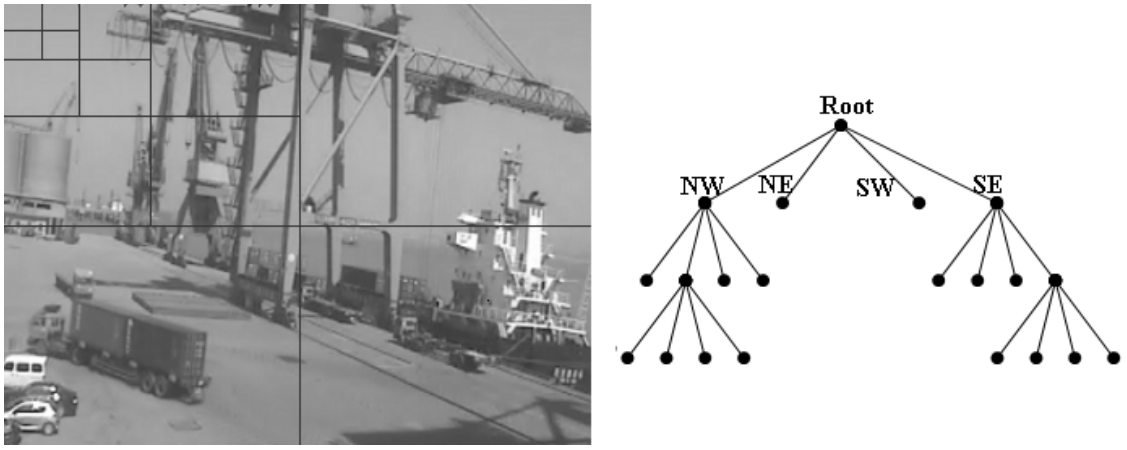

Fig.1.The hierarchical data structure of the quadtree

The hierarchical data structure of the quadtree[10] is shown in Fig. 1. The root node of a quadtree corresponds to the entire frame, and four children of one certain node correspond to equal-sized quadrants of the region represented by that node. The quadtree based on decomposition is applied to every input image of the video sequence. For each node at level L_{start}, a random pixel is sampled for the classification of a pixel as foreground or background where L_{start} is set by the user according to the system needs, in order to process the background subtraction algorithm at a certain processing speed. If the pixel is classified as foreground, then a randomly selected pixel in each of the four children nodes is sampled. If the pixel is classified as background, the next node at level L_{start} is considered. For each scanned node at level L_{final}, all pixels contained by the node are sampled, and if this node contains a large number of foreground pixels, its four-connected neighbors are scanned. By recursively repeating this process, fine boundaries of the foreground objects can be obtained. Table 1 shows the proposed algorithm in pseudo code.

Conventional methods, which raster scan entire image, cost O (D) time where D is the image size. Our hierarchical method just samples a small portion of the pixels in each image. Most of the background pixel is not sampled for foreground detection. The complexity is reduced to nearly O(R) where R is the moving foreground size. This idea is derived from [10], but our method is more efficient. In paper [10], MoG is adopted for pixel-based modeling, in which large portion of pixels is not sampled for classification; however parametric model updating, which can not be ignored, is particularly time-consuming. In our mean shift based non-parametric background modeling frame work, the model updating function Γ, which has to be implemented, is particularly simple.

TABLE I
COARSE TO FINE BACKGROUND SUBTRACTIONALGORITHM IN PSEUDOCODE

```
Main(){
    for(each node i at level L_start )
        Sampling(i);
    for(each sampled node i at level L_final)
      if(amount of foreground pixels in i >T)
        for( each four-connected node c)
            Sampling(c);
}

Sampling(node b){
    if(b.layer== L_final){
        for(each pixel x in block b)
            x.label=Φ(M(t),p);
        return;
      }
      p=randomPixel(b);
      p.label=Φ(M(t),p);
      if (p.label==foreground){
        Sampling(b.nw);
        Sampling(b.ne);
        Sampling(b.sw);
        Sampling(b.se);
      }
      else return;
}
```

IV. FAST MEAN SHIFT FOR SAMPLED PIXEL

Using quadtree based in the hierarchical structure that samples a small portion of the pixels in each frame, We can significantly reduce the redundant computational cost. However, for each sampled pixel, if the mean shift vector $m(x)$ is calculated via using the whole original data set, as shown in Equation (2), the computation load of mean shift iteration is still great. Furthermore it is very memory

intensive, since the past values of the pixel have to be recorded. By exploiting the fact that there is typically little change between successive video frames, we simplify the mean shift procedure in the following two aspects. Firstly, we perform mean shift iterations on incoming data and the mode found last time instead of the original data set. Secondly, the original data set is decomposed into a number of "local subsets", where there is very little change within local subsets. The samples of each local subset are treated as a whole in describing the density distribution, and assumed to come from the same class. The mean shift iterative procedure can be simplified furthermore to run on previous modes and such local subsets (shown as Fig. 2).

A. Partition step

As described in last subsection, a pixel is not sampled in each frame. For most of background blocks, only one random pixel is sampled for each frame, which means a pixel is sampled every K frames (*K* is block size), while the pixels often covered by moving objects are sampled frequently, K is consequently very small. Let $X=\{x_i\}_{i=1,2,\ldots,K}$ be a set of incoming data for a pixel between two sampling. We Divide X into m local subsets $S=\{s_i(c_i,n_i)\}_{i=1,2,\ldots,m}$ each with size n_i and center c_i. the partition step is described as follow:

I. Initialize S as $s_1(x_1,1)\}$, and s_1 is head node.

II. Then, for i=2 to *K* do the following.

III. Calculate the distances between x_i and head node(c_{head},n_{head}). Once if $\|x_k-c_{head}\|<T_r$, the head node is updated as Equation (6) and Equation (7), and go to the next iteration.

IV. If $\|x_k-c_{head}\|\geq T_r$, add as a subset (x_k,1) as head node

$$n_{head} = n_{head} + 1 \tag{6}$$

$$c_{head} = c_{head} + (x_k - c_{head})/ n_{head} \tag{7}$$

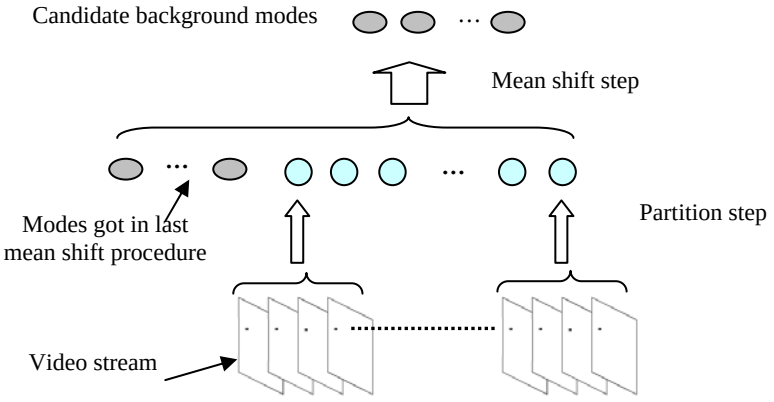

Fig.2. Fast mean shift for sampled pixel

B. Mean shift step

Let Y={(y_i , n_i^y)| i=1,…,Q} be the underlying density modes located in the last mean shift procedure, where y_i is the convergent point, and n_i^y is the number of points belonging to this mode. Let S={ $s_i(c_i,n_i^c) \mid i=1,...,M$ } be the local subsets with the cluster center c_i , and size n_i^c . The mean shift procedure is performed on $Z=\{(z_j,n_j^z) \mid j=1,...,M+Q\}$ which contains *Q* modes and *M* local subsets.

$$m(x)=\frac{\sum_{j=1}^{M+Q} n_j^z z_j exp(-\frac{1}{2}D^2(x,z_j;H))}{\sum_{j=1}^{M+Q} n_j^z exp(-\frac{1}{2}D^2(x,z_j;H))} \tag{8}$$

The convergent points are recorded as $\{\mu_k \mid k=1,...,K\}$. Let $z_{k_i,\ i=1,...,v}$ be the starting locations for which mean shift procedure converged to μ_k, , then n_k^μ , which is the number of frames for mode μ_k , can be calculated as Equation (9). Once if the candidate background modes are obtained, we can calculate the weight of each mode as Equation (4), and select the background models as Equation (5).

V. EXPERIMENTAL RESULTS

The proposed algorithm has been implemented on Pentium IV desktop with 3.0 GHz CPU and 1 GB of RAM. To evaluate the proposed background subtraction methods, four dynamic outdoor sequences are adopted as the benchmarks, including waving tree sequence with a person walking in front a heavily swaying tree, waving river sequence with ripples on the surface of the river, Trees and fountain sequence with trees swaying in the breeze, a fountain of water and undulating waves on the surface, and seaside scene sequence with , and the ocean waves ,which was taken from our marine video surveillance system The proposed method handles these situations robustly and the moving object is detected correctly. Fig. 3 shows the foreground detection results. The first column shows the input images, the second

column shows the foreground detection results by MoG, and the third column shows the foreground detection results by the proposed method. Morphological operators were not used in the results.

The speed of moving object is a critical factor that will affect many background models significantly. The average processing speeds of our method and MoG algorithms are presented in Table II. From the table, we can see that the computational complexity of the proposed background model is comparable to that of the MoG background model. The results are also evaluated quantitatively in terms of detection rate and false alarm rate by comparing to the manually segmented ground-truth foreground images. (see Table III)

$$Detection\ Rate = TP/(TP + FN) \quad (9)$$

$$False\ Alarm\ Rate = FP/(TP + FP) \quad (10)$$

Where *TP* is the number of the correctly detected moving object's pixels, *FN* is the missed moving object's pixels, and *FP* is the background pixels detected incorrectly as moving object's pixels.

The proposed algorithm has been applied to move objects detection in our real-time marine video surveillance system (Fig. 4). Based on streaming media via wide area network, the system detects and tracks moving objects on user-defined surveillance regions[8]. Whenever events, such as ships entering restricted zones, leaving port without permission, or moving in abnormal direction, is detected, alarms are generated and displayed to human operators for possible intervention.

Fig. 3. Foreground detection results

TABLE II
PROCESSING SPEED (FRAMES PER SECOND ON 3.0 GHZ CPU)

Video sequences	MOG	Our method
Waving trees	53.4	45.7
waving river	15.8	14.2
Trees and fountain	48.1	42.3
Seaside scene	23.7	25.3

TABLE III
QUANTITATIVE EVALUATIONS BY USING DETECTION RATE AND FALSE ALARM RATE

Video sequences	Detection Rate (%)		False Alarm Rate (%)	
	MOG	Our method	MOG	Our method
Waving trees	94.8	95.7	32.6	10.7
waving river	93.5	94.8	9.4	4.6
Trees and fountain	94.2	95.9	18.5	6.2
Seaside scene	92.7	92.6	8.7	0.4

Fig. 4. Real-time marine video surveillance system

VI. CONCLUSIONS AND FUTURE WORK

This paper presents mean shift based non-parametric background model for more sensitive and robust detection of moving objects in dynamic outdoor scenes. There are two main contributions in this paper. First, we proposed a coarse to fine method to segment the moving objects in the framework of nonparametric background model. Second, we presented a fast mean shift algorithm to reduce the complexity of background identification of a sampled pixel. Experiments on real videos show that the proposed background model generates better results than the well-known MoG background model with a bit more expensive computational cost.

ACKNOWLEDGEMENTS

The research is supported by the Key Project of the National Natural Science Foundation of China under Grant No.60533080.

REFERENCES

[1] A. Elgammal, D. *Harwood*, L. Davis, Non-parametric model for background subtraction, in: European Conference on Computer Vision, Dublin, Ireland, 2000, pp. 751-767.

[2] A. Mittal, N. Paragios, Motion-based background subtraction using adaptive kernel density estimation, IEEE International Conference on Computer Vision and Pattern Recogntion ,2004 , pp. 302–309

[3] B. Han, D. *Comaniciu*, and L. Davis, Sequential kernel density approximation through mode propagation: Applications to background modeling, Asian Conference on Computer Vision , 2004.

[4] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: realtime tracking of the human body, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) ,1998, pp.780-785.

[5] C. Stauffer, W. Grimson, Adaptive background mixture models for real-time tracking, IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 246-252.

[6] D.S. Lee, J.J. Hull, B. Erol, A Bayesian framework for Gaussian mixture background modeling. IEEE International Conference on Image Processing, 2003, pp.973-976.

[7] J. Han, P. Lan, X. Zhou, M. Zhang, "Automated visual surveillance System for port activity monitoring", Journal of Computational Information Systems, 1(2), 2006, pp.337-341.

[8] J. Han, Z. Pan, Robust Moving Objects Detection in Dynamic Scenes Based on Mean Shift, Proceedings of the international conference on Cyberworlds, 2008, pp. 271-275

[9] J. Wang, B. Thiesson, Y. Xu, and M. Cohen. Image and video segmentation by anisotropic mean shift, In Proceedings of European Conference on Computer Vision, pp.2004:238-249.

[10] J. Park, Hierarchical Data Structure for Real-Time Background Subtraction, IEEE Conference on Image Processing, 2006,pp.1849-1852.

[11] K. Zhang, M. Tang, J.T. Kwok, Applying neighborhood consistency for fast clustering and kernel density estimation, In Proceedings of Computer Vision and Pattern Recognition, 2007,pp.1001-1007 .

[12] M. Piccurdi, Z. Jan, Mean-shift background image modeling, IEEE International Conference on Image Processing, 2004,3399-3402. [12]

[13] O. Javed, K. Shafique, and M. Shah. A Hierarchical Approach to Robust Background Subtraction using Color and Gradient Information. IEEE Workshop on Motion and Video Computing, 2002, pp.22-27.

[14] R. Pless, , J. Larson, , S. Siebers, , B. Westover, Evaluation of local models of dynamic backgrounds. In: Computer Vision and Pattern Recognition, vol. II(2003), pp. 73-78

[15] Y. Chen, C. Chen, C. Huang, T. Hung, Efficient hierarchical method for background subtraction, Pattern Recognition, 40(2007), pp.2706-2715

[16] Y. Ke, R. Sukthankar, M. Hebert. Efficient Temporal Mean Shift for Activity Recognition in Video, NIPS Workshop on Activity Recognition and Discovery, 2005,124-127

[17] Y. Liu, H. Yao, W. Gao, X. Chen and D. Zhao, Nonparametric background generation, Journal of Visual Communication and Image Representation， 2007,18(3), pp.253-263.

[18] Y. Sheikh and M. Shah, Bayesian Object Detection in Dynamic Scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, San Diego,CA, 2005, pp. 74 -79

Jianping Han received the BS and MS degrees both in Computer Science from Nanjing University of Aeronautics & Astronautics in July 1990 and March 1996, respectively. Now, he serves as an associate professor of computer science at Hangzhou Dianzi University. His current research interests include computer vision, pattern recognition, and image processing.

Zhigeng Pan received his BS Degree and MS Degree from the Computer Science Department in 1987 and 1990 from Nanjing University respectively and Ph.D Degree in 1993 from Zhejiang University. Since 1993, he has been working at the State Key Lab of CAD&CG on a number of academic and industrial projects related with distributed graphics, virtual reality, and multimedia. He has published more than 70 papers on international journals, national journals and international conferences.

MingMin Zhang received her BS degree from Computer Science Dept, Nanjing University in 1990, and the MS degree from Computer Science and Engineering Department, Zhejiang University in 1995. She is now an associate professor of Computer and Engineering Department, Zhejiang University. Her research interests include virtual reality/virtual environment, multi-resolution modeling, real-time rendering, distributed VR, visualization, multimedia and image processing.